

CLUSTERIZACIÓN DE DATA PARA LA TOMA DE DECISIONES: ATENCIONES DE ASEGURADOS SIS EN EL PRIMER NIVEL DE ATENCIÓN EN LA RED DE SALUD AREQUIPA CAYLLOMA

Vladimir, Elker y Vidal

August 20, 2024

1 Introducción

La Red de Salud Arequipa Caylloma, como una de las 8 UEs de salud del Gobierno Regional de Arequipa, cumple con garantizar la gestión de los servicios de salud en el primer nivel de atención, en 147 establecimientos de salud, que abarcan la jurisdicción de las provincias de Arequipa y Caylloma en términos generales. [8] Además, el Gobierno Regional de Salud Arequipa, firma convenios con el Seguro Integral de Salud, para garantizar la gratuidad de las atenciones de los asegurados al SIS en todas las UEs de salud bajo el ámbito de su jurisdicción, lo que conlleva un financiamiento por parte del SIS que está condicionado al cumplimiento de los objetivos e indicadores establecidos. En la UE, la Oficina de Seguros se encarga de gestionar la información de las atenciones, las cuales son registradas a través del sistema ARFSIS Web (sistema local), luego enviados a los servidores del SIS, para que se consoliden en su base de datos central ubicada en la Sede Central del SIS - Lima.[4] La data que gestiona la Oficina de Seguros es la que se puede obtener de los reportes que genera el sistema SIGEPS (sistema web), que está conectado directamente con la base de datos central, teniendo información actualizada (en caliente). Sin embargo, la información que se genera en estos reportes es limitada. Aunque garantiza la inclusión de todas las atenciones SIS, solo muestra datos con respecto a datos del asegurado, establecimiento de salud, el tipo de servicio de salud, mas no información específica sobre parámetros como tipo de diagnóstico, procedimientos, etc, las cuales podrían servir para identificar tendencias específicas y mejorar la toma de decisiones.[7] Al respecto, la Oficina de Estadística e Informática, a través del sistema HIS (local), genera reportes semiestructurados más detallados, pero no diferencia de forma exacta si la atención fue Atención SIS o NO SIS. Ambas oficinas reportan la información de forma separada a la dirección de la UE, lo que genera una base de conocimiento incompleta para una adecuada toma de decisiones, por lo que es necesario implementar el análisis de datos que provengan de la Oficina de Seguros y la Oficina de Estadística e Informática, para una mejor gestión de la información y permitir un análisis adecuado para la toma de decisiones de nivel directoral.


| | | | | | |
|--|--|---|--|--|---|
|  PERÚ | | Ministerio de Salud | | Seguro Integral de Salud | |
| FORMATO UNICO DE ATENCION | | | | | |
| NUMERO DE FORMATO | | | | | |
| 220 - 12 - 00000001 | | | | | |
| CODIGO E.S./EQUIPO ASIST. 150509A201 | | NOMBRE DEL ESTABLECIMIENTO O EQUIPO ASIST. QUE REALIZA LA ATENCION C. S. MALA | | | RECONSIDERACION (*) <small>(*) Informar sobre la situación actual del establecimiento</small> |
| COMPONENTE SUBSIDIADO <input checked="" type="checkbox"/> SEMI-SUBSIDIADO <input type="checkbox"/> | TIPO FORMATO DE ATENCION NUEVO <input checked="" type="checkbox"/> ANTIGUO <input type="checkbox"/> | CODIGO AFILIACION / INSCRIPCION DATA: 220 2 NUMERO: 43639239 | | IDENTIFICACION TIPO: 1 N° DOCUMENTO: 43639239 | |
| FECHA DE NACIMIENTO DIA: 12 MES: 07 AÑO: 1986 | | SEXO MASCULINO <input type="checkbox"/> FEMENINO <input checked="" type="checkbox"/> | | ATENCION AMBULATORIA <input checked="" type="checkbox"/> REFERENCIA <input type="checkbox"/> EMERGENCIA <input type="checkbox"/> | |
| FECHA DE ATENCION DIA: 15 MES: 01 AÑO: 2012 | | HORA 11 : 15 | | LEGAR DE ATENCION INTRAMURAL <input checked="" type="checkbox"/> EXTRAMURAL <input type="checkbox"/> | |
| PERSONAL QUE ATIENDE DEL ESTABLECIMIENTO <input checked="" type="checkbox"/> ITINERANTE EQ. ASIST. <input type="checkbox"/> | | CODIGO DE PRESTACION 056 | | CODIGO E.S./ EQ. ASIST. 150509A201 | |
| DESTINO DEL ASSEGUADO ALTA <input checked="" type="checkbox"/> CITA <input type="checkbox"/> | | REEMBOLSO EMERGENCIA <input type="checkbox"/> CONSULTA EXTERNA <input type="checkbox"/> APOYO AL DIAGNOSTICO <input type="checkbox"/> | | CONTRARREEMBOLSO CONTRARREEMBOLSO <input type="checkbox"/> FALLECIDO <input type="checkbox"/> | |
| FECHA DE INGRESO DIA: 12 MES: 07 AÑO: 1986 | | FECHA DE ALTA DIA: 15 MES: 01 AÑO: 2012 | | FECHA DE PARTO DIA: 12 MES: 07 AÑO: 1986 | |
| CODIGO DEL E.S. 150509A201 | | E.S. AL QUE SE REFIERE CONTRARREEMBOLSO 150509A201 | | N° DIA DE REFERENCIA 1 | |

Figure 1: FUA(Formato Unico de Atención)

2 resumen

El análisis de la evolución espaciotemporal ha mostrado que la complejidad y gran escala de los datos, como los relacionados con la contaminación del aire, requieren de técnicas avanzadas como la clusterización para identificar patrones significativos. De manera similar, en el contexto de las atenciones de salud, los datos generados a partir de múltiples fuentes (por ejemplo, datos de atenciones SIS y NO SIS en los establecimientos de salud) también presentan una gran dimensionalidad y diversidad temporal y espacial. La clusterización se vuelve esencial en este contexto para agrupar datos con características similares y detectar patrones emergentes que podrían no ser evidentes con métodos de análisis más simples.[7] Al aplicar técnicas de clusterización, como K-means, se pueden identificar subgrupos de atenciones que comparten tendencias comunes, lo que permite a los responsables de la toma de decisiones enfocarse en

áreas críticas y asignar recursos de manera más efectiva.[11] Además, la visualización de estos clusters en gráficos de dispersión y mapas facilita la comprensión de la distribución geográfica y temporal de las atenciones, lo que es crucial para mejorar la gestión y planificación de los servicios de salud. En resumen, la clusterización no solo optimiza la interpretación de grandes volúmenes de datos, sino que también apoya la toma de decisiones informadas y orientadas a resultados en el ámbito de la salud

3 abstract

The analysis of spatio-temporal evolution demonstrates that the complexity and large scale of data, such as those related to air pollution, necessitate advanced techniques like clustering to identify meaningful patterns. Similarly, in the context of healthcare data, such as those generated from various sources (e.g., SIS and non-SIS healthcare services),

the data also exhibit high dimensionality and temporal and spatial diversity. Clustering becomes essential in this scenario for grouping data with similar characteristics and detecting emerging patterns that might not be apparent through simpler analysis methods. By applying clustering techniques like K-means, it is possible to identify subgroups of healthcare services that share common trends, allowing decision-makers to focus on critical areas and allocate resources more effectively. Additionally, visualizing these clusters in scatter plots and maps facilitates the understanding of the geographic and temporal distribution of healthcare services, which is crucial for improving the management and planning of health services. In summary, clustering not only optimizes the interpretation of large volumes of data but also supports informed and result-oriented decision-making in the healthcare sector.

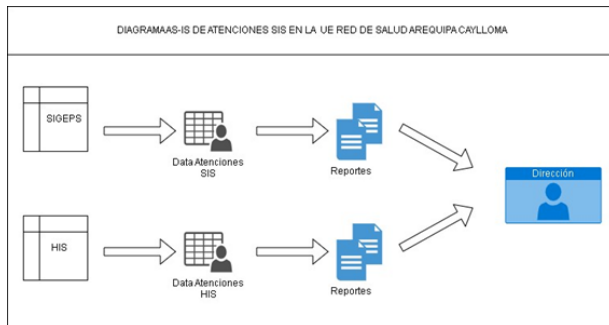


Figure 2: Diagrama AS-IS de Gestión de Atenciones SIS en la Red de Salud Arequipa Caylloma

4 keyword

Spatio-temporal evolution, clustering, K-means, healthcare data, decision-making, pattern detection, data analysis, air pollution, sequence mining, dynamic time warping, visual analytics, geographic distribution.

5 Problema

El problema a resolver es la falta de consolidación y procesamiento de la información de las atenciones SIS en la RSAC, lo que impide a la Dirección de la UE acceder a datos relevantes y actualizados para la toma de decisiones estratégicas. Esta deficiencia en la gestión de la información se traduce en una limitada capacidad para garantizar la calidad de las atenciones, optimizar los recursos, y responden de manera efectiva a las necesidades de los pacientes.

6 trabajos relacionados

La aplicación de modelos predictivos en la salud pública ha ganado importancia, especialmente en la predicción de enfermedades como la anemia infantil. Estos modelos permiten identificar factores de riesgo y predecir la probabilidad de que un niño desarrolle anemia, utilizando variables que incluyen datos demográficos, nutricionales y otros factores clínicos. En el estudio realizado por Valdez et al. (2023), se emplearon técnicas de minería de datos para predecir la anemia en niños menores de cinco años en el Perú, utilizando un conjunto de datos extraído de la plataforma de datos abiertos del gobierno peruano. En este estudio se aplicaron varios algoritmos de aprendizaje automático, como Naive Bayes, árboles de decisión, regresión logística, K vecinos más cercanos y bosques aleatorios. Los resultados mostraron que el algoritmo de Naive Bayes fue el más efectivo, con un recall del 74% y una precisión del 43%, lo que sugiere que este enfoque es particularmente adecuado para conjuntos de datos desequilibrados (Valdez et al., 2023).[2] Estudios similares han aplicado modelos predictivos en diferentes contextos geográficos. Por ejemplo, en Afganistán, Momand et al. (2020) utilizaron clasificadores

como Random Forest y Naive Bayes para predecir la desnutrición en niños, logrando una precisión superior al 90%. De manera similar, Ferreira et al. (2018) emplearon técnicas de minería de datos en Portugal para predecir la necesidad de intervención nutricional en pacientes, utilizando una combinación de clasificadores y evaluando su rendimiento mediante medidas como la precisión y la tasa de error. Estos estudios subrayan la importancia de seleccionar el modelo adecuado y realizar un preprocesamiento exhaustivo de los datos para obtener resultados efectivos en la predicción de enfermedades. En el caso de la anemia infantil, la capacidad de identificar a los niños en riesgo permite a los responsables de salud pública tomar medidas preventivas y dirigir recursos de manera más eficiente, contribuyendo así a la mejora de la salud infantil en comunidades vulnerables. La comparación de resultados entre diferentes estudios destaca la necesidad de adaptar los modelos a las características específicas de los datos y el contexto en el que se aplican.[1] En resumen, la literatura existente sugiere que los modelos de aprendizaje automático, cuando se aplican correctamente, pueden ser herramientas poderosas para la detección temprana de enfermedades como la anemia, proporcionando información valiosa para la toma de decisiones en salud pública [10].

7 Marco Teórico

7.1 Minería de Datos en la Salud

La minería de datos es un proceso crucial en la extracción de patrones útiles y conocimiento a partir de grandes volúmenes de datos. En el contexto de la salud, esta disciplina ha adquirido relevancia debido a la creciente cantidad de datos generados por

los sistemas de información en salud, como los registros electrónicos de salud, bases de datos de asegurados, y sistemas de monitoreo de enfermedades. La minería de datos permite a los investigadores y responsables de la toma de decisiones en salud descubrir patrones ocultos, correlaciones significativas, y tendencias emergentes, facilitando así la identificación de áreas críticas para la intervención y la mejora en la calidad de los servicios de salud.

7.2 Clusterización y K-means

La clusterización es una técnica de análisis exploratorio de datos que tiene como objetivo agrupar un conjunto de objetos en subgrupos o clusters de modo que los objetos dentro de un mismo grupo sean más similares entre sí que con los de otros grupos. El algoritmo K-means es uno de los métodos de clusterización más utilizados debido a su simplicidad y eficiencia. K-means agrupa los datos en K clusters, minimizando la variación dentro de cada cluster. En la gestión de salud, la clusterización con K-means puede ayudar a identificar patrones de atención, segmentar a los pacientes en grupos con características similares, y detectar áreas con altos índices de demanda de servicios

7.3 Análisis Espaciotemporal

El análisis espaciotemporal es una técnica que examina la evolución de fenómenos a través del tiempo y el espacio. En salud pública, el análisis espaciotemporal se utiliza para identificar la propagación de enfermedades, evaluar la efectividad de intervenciones sanitarias y estudiar la distribución geográfica de las atenciones de salud. La integración de técnicas de clusterización con el análisis espaciotemporal permite una mejor comprensión de cómo las tendencias en la atención médica varían según la localización y el tiempo, pro-

porcionando una base sólida para la planificación de recursos y la toma de decisiones

7.4 Gestión de la Información en Salud

La gestión de la información en salud implica la recolección, almacenamiento, análisis y uso de datos para mejorar los resultados en salud. En el contexto de la Red de Salud Arequipa Caylloma, la gestión efectiva de la información es esencial para garantizar que las decisiones se basen en datos precisos y actualizados. El Sistema ARFSIS Web y el sistema HIS son herramientas fundamentales en la recopilación de datos de atenciones SIS y no SIS, aunque presentan limitaciones en la integración y análisis conjunto de la información. La implementación de técnicas avanzadas de minería de datos y clusterización permite superar estas limitaciones, ofreciendo una visión más completa y detallada del estado de salud de la población.

7.5 Visualización de Datos

La visualización de datos es un componente esencial en la minería de datos y el análisis espaciotemporal, ya que permite a los analistas y tomadores de decisiones interpretar grandes volúmenes de datos de manera intuitiva y rápida. Gráficos de dispersión, mapas y otras técnicas de visualización son herramientas efectivas para representar la distribución geográfica y temporal de los clusters identificados, facilitando la identificación de patrones y tendencias clave que podrían no ser evidentes en análisis tabulares o lineales. La visualización mejora la capacidad de respuesta de los sistemas de salud al proporcionar una comprensión clara y accesible de los datos complejos

7.6 Aplicaciones y Relevancia del Clustering en la Salud

El clustering ha demostrado ser una herramienta valiosa en diversos estudios de salud pública y medicina. Por ejemplo, ha sido utilizado para la segmentación de pacientes en grupos de riesgo, la identificación de brotes de enfermedades infecciosas, y la optimización de recursos en hospitales. La relevancia del clustering en la salud radica en su capacidad para sintetizar grandes volúmenes de datos en grupos manejables, lo que permite a los profesionales de la salud focalizar sus esfuerzos en intervenciones más precisas y efectivas.

8 Analisis de Tareas

La propuesta estará enmarcada desde un enfoque de mejora continua; es decir, el proceso será cíclico, luego de culminado el análisis de resultados para la toma de decisiones de nivel directoral, se volverá a iniciar para identificar nuevos datos y por consiguiente generar nuevo conocimiento.

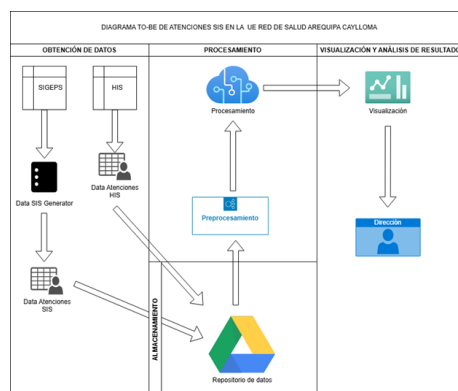


Figure 3: Diagrama TO-BE de atenciones SIS en la UE red de salud Arequipa - Caylloma [5]

8.1 Obtención de Datos

La Data HIS se genera directamente del sistema HIS-MINSA, a través de reportes men-

suales generados en formato EXCEL, a nivel de UE (RSAC). La Data SIS se genera del sistema del SIS, SIGEPS, a través de reportes mensuales generados en formato EXCEL, a nivel de establecimiento de salud (IPRESS). Para facilitar una obtención de datos de forma oportuna, se implementó un script que permite realizar web scrapping al sistema, permitiendo la generación de los reportes mensuales correspondientes a las 147 IPRESS de la jurisdicción de la UE (RSAC), de forma automatizada, consolidando toda la data a nivel de UE. [9]

Adicionalmente, se vió la necesidad de hacer uso del dataset CIE-10, el cual contiene el listado de códigos de diagnósticos de la Clasificación Internacional de Enfermedades, Decima Edición, la cual se obtuvo de datos públicos proporcionados por el Ministerio de Salud.

Además, se incluyó un dataset IPRESS, el cual contiene el listado de los 147 establecimientos de salud (IPRESS) del ámbito de la jurisdicción de la Red de Salud Arequipa Caylloma, con sus respectivos códigos de IPRESS que se maneja a nivel de HIS y SIS, lo cual permitirá identificar los establecimientos en posteriores análisis.

8.2 Almacenamiento

La data obtenida se almacena en un repositorio en la nube (Google Drive) para que se encuentre disponible de forma oportuna para un posterior procesamiento de datos

The image displays three screenshots of a Google Drive interface, showing the hierarchy of folders and files for a project named 'Data'.

Screenshot 1: 'Data' folder view

The breadcrumb path is 'Mi unidad > Recuperacion Informa... > Data'. The view shows a list of folders:

| Nombre | Propietario |
|--------------------|-------------|
| _ANÁLISIS | yo |
| _PRE-PROCESAMIENTO | yo |
| CIE-10 | yo |
| HIS | yo |
| IPRESS | yo |
| SIS | yo |

Screenshot 2: 'Data > SIS' folder view

The breadcrumb path is '... > Data > SIS'. The view shows a list of Excel files:

| Nombre | Propietario |
|---------------------------------|-------------|
| 2024_01_PROFESIONALES_EESS.xlsx | yo |
| 2024_02_PROFESIONALES_EESS.xlsx | yo |
| 2024_03_PROFESIONALES_EESS.xlsx | yo |
| 2024_04_PROFESIONALES_EESS.xlsx | yo |

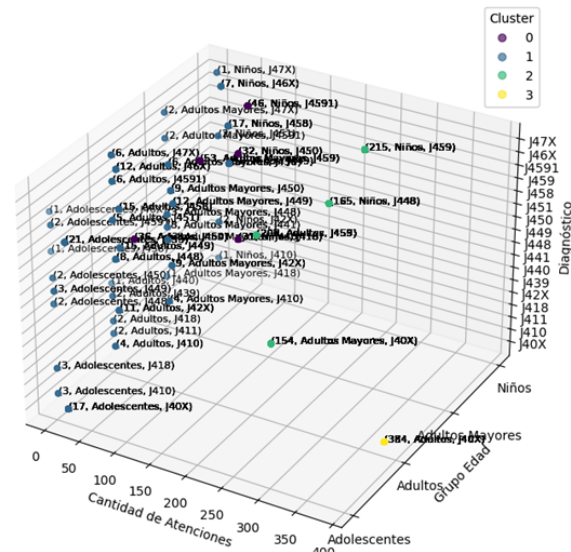
Screenshot 3: 'Data > HIS' folder view

The breadcrumb path is '... > Data > HIS'. The view shows a list of Excel files:

| Nombre | Propietario |
|-----------------|-------------|
| 01.ENERO.xlsx | yo |
| 02.FEBRERO.xlsx | yo |
| 03.MARZO.xlsx | yo |
| 04.ABRIL.xlsx | yo |



Análisis de Clusters en 3D: Cantidad de Atenciones, Grupo Edad y Diagnóstico



8.3 Procesamiento

8.3.1 pre-procesamiento



Figure 4: Data Pre- Procesada

8.3.2 procesamiento

Análisis de atenciones en base a edad y enfermedades respiratorias crónicas través de clustización utilizando algoritmo K-MEANS

8.4 Visualización

Visualización de información K-MEANS: edad y diagnóstico de enfermedades crónicas

El gráfico generado con K-means en 3D, que agrupa las observaciones en diferentes clusters basados en la Cantidad de Atenciones, Grupo de Edad, y Diagnóstico, ofrece varias conclusiones potenciales que podrían ser extraídas del análisis:

1. Identificación de Grupos de Pacientes con Patrones Similares:
 - El uso de K-means permite agrupar a los pacientes en clusters según características comunes. Por ejemplo, los pacientes en el Cluster 0 (Morado) pueden compartir similitudes en términos de cantidad de atenciones y diagnósticos dentro de sus respectivos grupos de edad. Este grupo representa el mayor número de casos y podría indicar un patrón prevalente en el tipo de enfermedades o condiciones tratadas.
 - Cluster 1 (Verde), Cluster 2 (Celeste), y Cluster 3 (Amarillo) representan subgrupos más pequeños con características específicas. Por ejemplo, Cluster 3 (Amarillo), que incluye a adultos mayores con un diagnóstico particular y una alta cantidad de atenciones, podría indicar una necesidad de atención especial o más recursos para estos pacientes.

2. **Identificación de Diagnósticos Críticos:** La concentración de puntos en ciertos clusters sugiere que ciertos diagnósticos son más comunes dentro de grupos específicos de edad. Si un diagnóstico particular se agrupa en un cluster con alta cantidad de atenciones, como en Cluster 0 (Morado), esto puede indicar que estas enfermedades requieren un enfoque preventivo o más recursos en términos de atención médica. Diagnósticos en clusters más pequeños o más alejados (como en el Cluster 3 (Amarillo)) podrían representar condiciones menos comunes pero más severas, que también requieren atención especial.
3. **Diferencias en la Distribución de Atenciones por Grupo de Edad:** La distribución de los puntos en el gráfico sugiere que ciertos grupos de edad tienen una mayor o menor cantidad de atenciones para diagnósticos específicos. Por ejemplo, si los adultos mayores están mayormente agrupados en un cluster con altas atenciones, podría indicar un mayor uso de servicios de salud para condiciones crónicas en este grupo. Los Niños y Adolescentes podrían estar agrupados en clusters con diagnósticos relacionados a condiciones comunes en estas edades, como enfermedades respiratorias.
4. **Segmentación para Intervenciones Específicas:** Al identificar los clusters, se pueden diseñar intervenciones específicas para cada grupo. Por ejemplo, un cluster con Niños y Adolescentes que comparten un diagnóstico específico podría beneficiarse de programas de prevención en escuelas o campañas de vacunación específicas. Adultos Mayores en un cluster que muestra una alta cantidad de atenciones para diagnósticos crónicos podrían beneficiarse de un seguimiento más intensivo o programas de manejo de enfermedades crónicas.
5. **Priorización de Recursos de Salud:** Los clusters más grandes indican dónde se concentra la mayor carga de atención médica, lo que puede ayudar a priorizar recursos y esfuerzos en esos grupos. Por ejemplo, si un cluster grande está relacionado con enfermedades respiratorias en adultos, los recursos podrían dirigirse a fortalecer los servicios en esa área. Los clusters más pequeños pero críticos (como el Cluster 3 (Amarillo)) podrían indicar la necesidad de recursos adicionales o atención especializada para manejar condiciones menos comunes pero graves.

9 Minería de Datos con K-means

9.1 Exploración de la relación entre edad y atenciones

En esta exploración se considera la posibilidad de que exista una relación entre la edad de los pacientes y el número de atenciones recibidas, donde se hipotetiza que las edades extremas, como de 0 a 5 años y mayores de 60 años, recurren con mayor frecuencia a los centros de salud.

Para validar esta hipótesis, primero debemos realizar una agrupación por la columna **DOCUMENTO** (que corresponde a la identidad única de cada paciente), agregar una columna llamada **ATENCIONES** y eliminar los duplicados según la columna **DOCUMENTO**. Para este objetivo recurrimos a la siguiente técnica:

```
import pandas as pd
```



```

3 # Leer el archivo CSV
4 df = pd.read_csv('/content/drive/
  MyDrive/retrival_information/
  Trabajo Final/
  DATA_SIS_Consolidada_Preproceso.
  csv')
5
6 # Contar el n mero de atenciones
  por cada DOCUMENTO
7 atenciones_count = df.groupby('
  DOCUMENTO').size().reset_index(
  name='ATENCIONES')
8
9 # Unir el conteo de atenciones al
  DataFrame original
10 df = df.merge(atenciones_count, on='
  DOCUMENTO')
11
12 # Eliminar duplicados y mantener
  solo el primer registro para cada
  DOCUMENTO
13 df_unique = df.drop_duplicates(
  subset='DOCUMENTO', keep='first')
14
15 # Mostrar el DataFrame resultante
16 df_unique.head()

```

9.2 Cálculo del número de centroides utilizando el método del codo

Este método sugiere ejecutar K-means para un número de centroides que varía entre 2 y 12, y calcular la Suma de Cuadrados Dentro de los Clústeres (WCSS, por sus siglas en inglés). WCSS se refiere a la suma de las distancias cuadradas de cada punto de datos al centroide de su clúster. Para cada punto de datos, se calcula la distancia al centroide del clúster, se eleva al cuadrado, y luego se suman todas estas distancias cuadradas para todos los puntos en todos los clústeres.

```

1
2
3 # Leer el archivo CSV
4 df = pd.read_csv('/content/drive/
  MyDrive/retrival_information/
  Trabajo Final/
  DATA_SIS_Consolidada_Preproceso.

```

```

  csv')
5
6 # Seleccionar las columnas
  necesarias
7 data = df[['EDAD', 'ATENCIONES']]
8
9 # Normalizar los datos
10 scaler = StandardScaler()
11 data_scaled = scaler.fit_transform(
  data)
12
13 # Lista para almacenar el Total WCSS
  (Within-Cluster Sum of Squares)
  para cada valor de K
14 wcss = []
15
16 # Ejecutar K-means para K de 2 a 10
17 for k in range(2, 11):
18     kmeans = KMeans(n_clusters=k,
19                     random_state=0)
20     kmeans.fit(data_scaled)
21     wcss.append(kmeans.inertia_) #
22     Total WCSS para el modelo actual
23
24 # Plotear los resultados
25 plt.figure(figsize=(10, 6))
26 plt.plot(range(2, 11), wcss, marker=
27     'o', linestyle='--', color='b')
28 plt.title('Elbow Method for Optimal
29     K')
30 plt.xlabel('N mero de Cl steres (K
31     )')
32 plt.ylabel('')
33 plt.xticks(range(2, 11))
34 plt.grid(True)
35 plt.show()

```

Una vez realizado el cálculo, procedemos a visualizar la información. En el eje X se muestra el número de clústeres (de 2 a 12) y en el eje Y, el WCSS. El método sugiere que el número óptimo de centroides corresponde al punto donde se forma el "codo" o esquina en la gráfica.

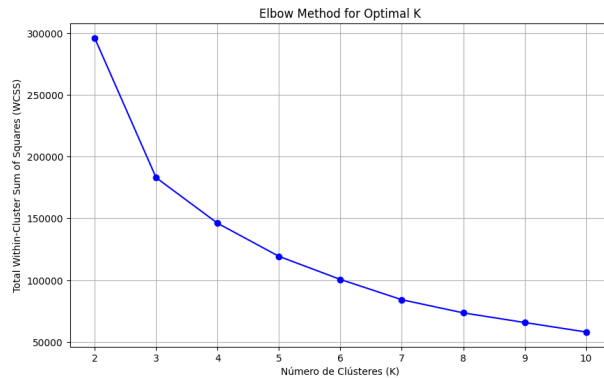


Figure 5: Elección del mejor numero de cluster

Según la gráfica obtenida, seleccionamos como óptimo un número de 3 clústeres.

9.3 Ejecución de K-means para 3 centroides

Ejecutamos el algoritmo K-means de la librería `sklearn` para un número de clústeres $k = 3$, utilizando el DataFrame con las columnas `EDAD` y `ATENCIONES`.

```

1 # Leer el archivo CSV
2 #df = pd.read_csv('/content/drive/
  MyDrive/retrival_information/
  Trabajo Final/
  DATA_SIS_Consolidada_Preproceso.
  csv')
3
4 # Seleccionar las columnas
  necesarias
5 data = df[['EDAD', 'ATENCIONES']]
6
7 # Normalizar los datos
8 scaler = StandardScaler()
9 data_scaled = scaler.fit_transform(
  data)
10
11 # Ejecutar K-means con 3 clústeres
12 kmeans = KMeans(n_clusters=3,
  random_state=0)
13 kmeans.fit(data_scaled)
14 clusters = kmeans.predict(
  data_scaled)
15
16 # Aadir la columna de clústeres
  al DataFrame original

```

```

17 df['Cluster'] = clusters
18
19 # Graficar los puntos con colores
  diferentes para cada clúster
20 plt.figure(figsize=(10, 6))
21 scatter = plt.scatter(df['EDAD'], df
  ['ATENCIONES'], c=df['Cluster'],
  cmap='viridis', marker='o')
22 plt.title('Clustering con K-means (3
  Clústeres)')
23 plt.xlabel('Edad')
24 plt.ylabel('Atenciones')
25 plt.colorbar(scatter, label='Número
  de Clúster')
26 plt.grid(True)
27 plt.show()

```

9.4 Interpretación de la gráfica

La gráfica presentada muestra los resultados de un análisis de agrupamiento utilizando el algoritmo K-means, considerando dos variables: **edad** y **número de atenciones** de los pacientes. A continuación, se ofrece una interpretación detallada.

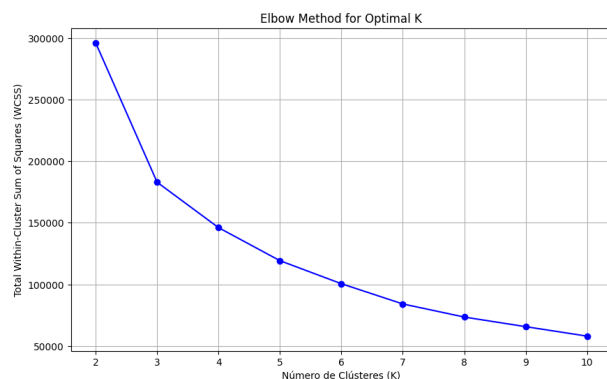


Figure 6: Elección del mejor numero de cluster

Ejes

- El eje **X** representa la **edad** de los pacientes.
- El eje **Y** representa el **número de atenciones** que han recibido los pacientes.

Número de Clústeres

- Se han formado **3 clústeres** diferentes, los cuales están codificados por colores.
- Cada clúster agrupa a pacientes que presentan características similares en términos de edad y número de atenciones.

Interpretación de los Clústeres

- **Clúster 0 (color amarillo):** Representa principalmente a pacientes jóvenes (aproximadamente entre 0 y 20 años) que han recibido un número bajo de atenciones (aproximadamente menos de 10).
- **Clúster 1 (color verde azulado):** Este clúster incluye a pacientes de mayor edad, generalmente entre 40 y 100 años, con un número bajo a medio de atenciones.
- **Clúster 2 (color morado):** Agrupa a pacientes de todas las edades que han recibido un número considerablemente mayor de atenciones. Sin embargo, este clúster parece estar más concentrado en personas de mediana edad (20-60 años).

Observaciones Adicionales

- Se observan algunos puntos atípicos o *outliers* en la gráfica, como pacientes jóvenes con un número muy alto de atenciones, lo cual es inusual según el patrón general.
- El clúster 2 muestra una mayor dispersión en cuanto a la edad, lo que sugiere que existe una variedad de edades entre los pacientes que requieren más atenciones.

10 Área de interés

El área de interés de este proyecto se centra en la gestión y análisis de datos de atenciones de salud en los establecimientos del Primer Nivel de Atención (I-1, I-2, I-3, y I-4) de la Red de Salud Arequipa Caylloma (RSAC), particularmente en las atenciones financiadas por el Seguro Integral de Salud (SIS).[6] Este interés se enmarca en la necesidad de mejorar la calidad y efectividad de la toma de decisiones en la gestión de salud a nivel directoral en la RSAC

11 Tópico

El tópico principal del proyecto es la mejora de la gestión y el análisis de datos de atenciones de salud financiadas por el SIS en la Red de Salud Arequipa Caylloma. Esto incluye la integración y consolidación de datos provenientes de diferentes fuentes dentro de la unidad ejecutora (UE), como la Oficina de Seguros y la Oficina de Estadística e Informática, utilizando sistemas como SIGEPS (Oficina de Seguros) y HIS-MINSA (Oficina de Estadística e Informática).

12 Tema

El tema específico a abordar es la implementación del análisis exploratorio de datos de atenciones de salud que permita consolidar y procesar la información de las atenciones SIS para facilitar la toma de decisiones en la Dirección de la UE.[3] Actualmente, la información disponible a través del sistema SIGEPS no está suficientemente consolidada ni procesada, lo que limita la capacidad de los directivos de la RSAC para tomar decisiones informadas.

13 Artículo de referencia

Para el presente proyecto se toma como referencia el artículo **"Aprendizaje automático para predicción de anemia en niños menores de 5 años mediante el análisis de su estado de nutrición usando minería de datos"**, el cual proporciona un respaldo indirecto en el planteamiento de la propuesta, con respecto a la importancia de procesos como la limpieza de datos, la selección de características relevantes, y la aplicación de modelos predictivos para mejorar la comprensión y gestión de los datos en salud [1]

14 Variables Analizadas

1. NRO FORMATO: Número de Formato Único de Atención (FUA). Documento que se genera para la atención del afiliado SIS.
2. F. ATENCION: Fecha de atención. Indica el día en que se realizó la atención médica al beneficiario.
3. TIP. DOC.: Tipo de documento. Especifica el tipo de documento de identidad del beneficiario, como DNI, carnet de identidad.
4. DOCUMENTO: Número de documento de identidad del beneficiario. Este es el identificador único del paciente dentro del sistema.
5. CONTRATO: Número o código del contrato de afiliación al Seguro Integral de Salud (SIS).
6. BENEFICIARIO: Nombre del beneficiario, es decir, la persona que recibe la atención médica.
7. F. NACIMIENTO: Fecha de nacimiento del beneficiario. Este dato es importante para calcular la edad del paciente y evaluar el contexto de la atención.
8. EDAD: Edad del beneficiario al momento de la atención. Este campo se calcula a partir de la fecha de nacimiento y la fecha de atención.
9. SEXO: Sexo del beneficiario, representado como 'M' para masculino o 'F' para femenino.
10. EESS CODIGO: Código del Establecimiento de Salud (EES). Este código identifica el lugar donde se brindó la atención.
11. EESS NOMBRE: Nombre del Establecimiento de Salud del primer nivel de atención donde se realizó la atención. Es el nombre del centro de salud o puesto de salud.
12. SERVICIO: Código o descripción del servicio médico proporcionado, como consulta externa, hospitalización, emergencia, etc.
13. DNI PROFESIONAL: Número de DNI del profesional de salud que atendió al paciente. Es el identificador único del médico o personal de salud.
14. NOMBRE PROFESIONAL: Nombre del profesional de salud que brindó la atención.
15. TIPO PROFESIONAL: Tipo de profesional de salud (médico, enfermero, obstetra, etc.).
16. TARIFA: Tarifa o costo asociado al servicio brindado, que puede variar según el tipo de atención o contrato. Dato depreciable porque la información que se genera no es real.

17. HIST. CLINICA: Número de la historia clínica del paciente. Es un identificador interno del paciente en el sistema del Establecimiento de Salud.
18. COMPONENTE: Componente del servicio o atención brindada.
19. COND. MATERNA (*): Condición materna. Este campo se refiere a la condición de la madre en el caso de atenciones perinatales, donde se podría registrar si hay algún riesgo o complicación durante el embarazo o parto.
20. TIP. ATENCION (**): Tipo de atención recibida, como ambulatoria, hospitalización, urgencias, entre otros. Define la naturaleza del servicio brindado.
21. LUG. ATENCION (**): Lugar de atención, que puede ser dentro del mismo Establecimiento de Salud o en otro lugar, como a domicilio o en otra entidad.
22. EESS REFERENCIA: Establecimiento de salud de referencia. Indica si el paciente fue derivado a otro centro de salud para continuar con la atención o recibir un tratamiento específico.
23. F. REGISTRO: Fecha de registro de la atención en el sistema. Esta fecha puede diferir de la fecha de atención y se refiere a cuándo se ingresó la información en el sistema.
24. DIGITADOR: Nombre o código del digitador que ingresó la información al sistema.
25. NRO CRED: Número de controles de crecimiento y desarrollo.
26. MES: Mes correspondiente al periodo de digitación.

15 Analisis estadístico de datos

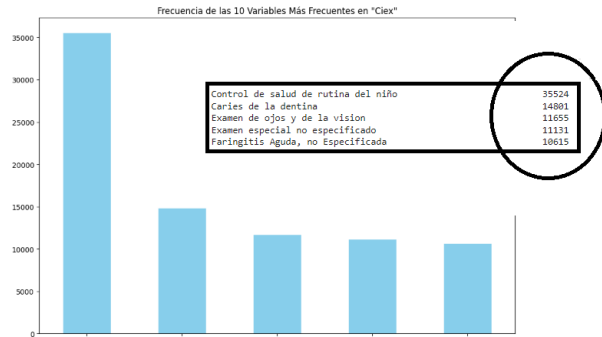
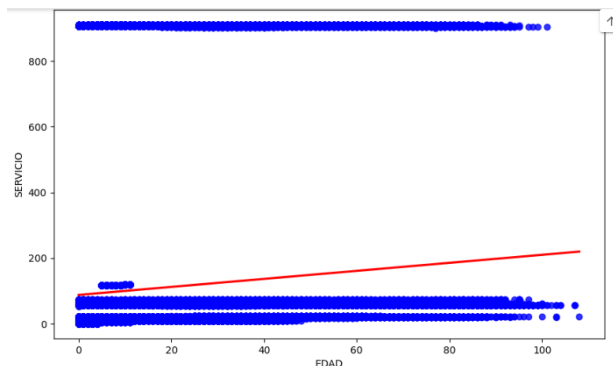


Figure 7: Motivos de frecuencia al centro de salud

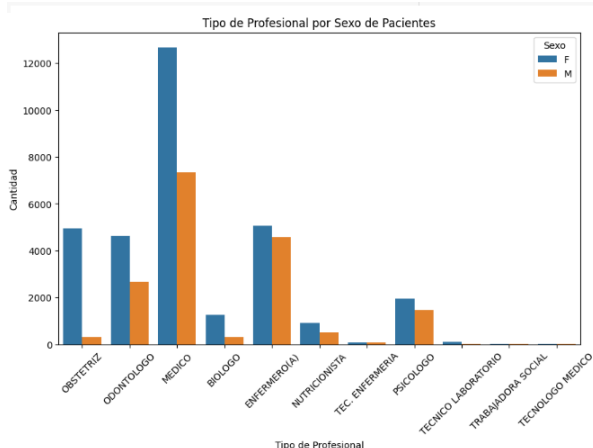
Observamos que la causa mas frecuente por la que se visita al centro de salud es por **rutina del niño** lo que nos lleva a plantear la hipótesis que los niños son los más propensos a contraer enfermedades. También existen factores, pues los padres tienden a llevar a sus hijos al centro de salud incluso por motivos que normalmente no requerirían atención médica. notamos que las variables EDAD y SERVICIO son variables del tipo cuantitativo y por lo tanto podemos analizar si existe alguna relación entre las variables. El grado de correlación de Pearson nos da un valor de 0.116577370, lo que indica una correlación débil entre las variables. Esto sugiere que hay una relación leve entre las dos variables, pero no es fuerte. Signo (+): Como el valor es positivo, significa que cuando la variable EDAD aumenta, la variable SERVICIO tiende a aumentar ligeramente también, pero esta relación es débil y puede no ser significativa.

Conclusión: El coeficiente de 0.1166 indica que la relación entre EDAD y SERVICIO es débil, lo que implica que los cambios en la edad de los sujetos no están fuertemente asociados con cambios en el servicio que reciben.

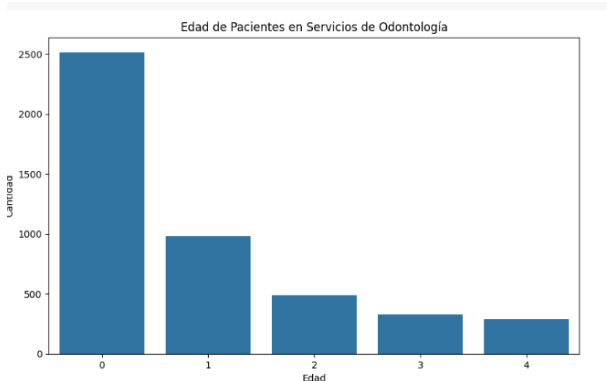
En términos prácticos, es posible que la edad no sea un factor determinante en el servicio.



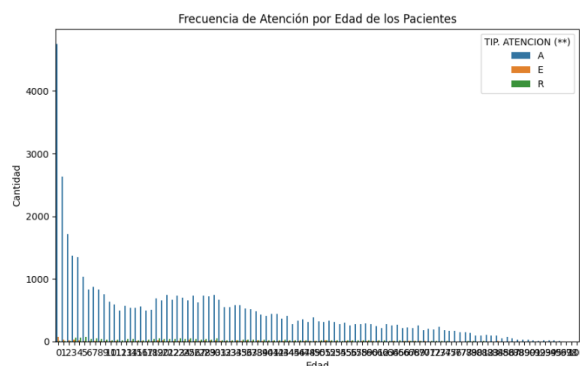
Podemos analizar también el número de personas que van a los hospitales dependiendo del sexo.



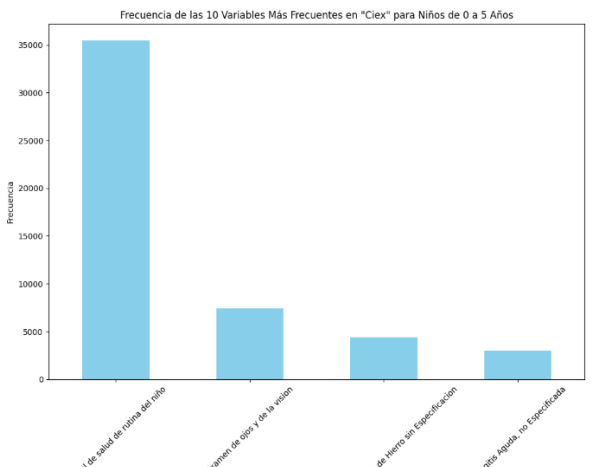
Vemos que en general el número de mujeres que se atiende en los hospitales supera a los hombres. Además las consultas médicas son los servicios que mayormente utilizan. De igual manera ocurre con el número de niños que frecuentan al dentista es mayor que los adultos.



En general la tasa de niños que asisten al médico es mayor que la de los adultos.



Al analizar los motivos por los que los niños asisten al centro de salud con mayor frecuencia, obtuvimos que los niños aparte de su control de rutina, asisten al médico para exámenes de visión y Anemia por deficiencia de hierro.



References

- [1] M. R. Ferreira, M. R. Pereira, and R. M. de Sousa. “Data mining techniques applied to the prediction of nutritional status”. In: *Journal of Biomedical Informatics* 85 (2018), pp. 56–65.
- [2] Karen Hayme Garcia Ortiz and Sandra Leandres Quispe. “Carga laboral y satisfacción de las enfermeras del servicio de emergencia del Hospital Nacional Carlos Alberto Seguin Escobedo-EsSalud, Arequipa-2017”. In: (2018).
- [3] Magali Latorre Delgado and Nelly Elvira Suclla Muñoz. “Percepción del asegurado sobre la calidad de atención en el servicio de emergencia del Hospital Base Carlos Alberto Seguin Escobedo, Essalud, Arequipa, 2016.” In: (2016).
- [4] Arturo Recabarren Lozada and Sandra Cárdenas Hilasaca. “Factores de riesgo de asma infantil en niños que asisten al Programa de Control de Asma del Hospital III Yanahuara Essalud-Arequipa”. In: *Enfermedades del Tórax* 46.2 (2003), pp. 118–125.
- [5] O. S. Momand, N. M. Malyar, and K. Kakar. “Predicting child malnutrition using machine learning methods: A case study in Afghanistan”. In: *International Journal of Medical Informatics* 140 (2020), p. 104143.
- [6] Jeanette Geraldine Núñez Borda. “Prevalencia, Características Clínicas e Histopatológicas del Carcinoma Basocelular en Pacientes Tratados en el Servicio de Dermatología del Hospital Base Carlos Alberto Seguin Escobedo Essalud-Arequipa de Enero del 2008 a Diciembre del 2018”. In: (2019).
- [7] Arturo Felipe Recabarren Lozada, Karen Yaneth Portugal Valdivia, and Javier Herbert Gutierrez Morales. “Comparación de las características clínicas del asma bronquial entre niños con sobrepeso/obesidad y niños eutróficos inscritos en el Programa de Asma Bronquial del Hospital III Yanahuara EsSalud-Arequipa”. In: *Diagnóstico (Perú)* (2003), pp. 60–67.
- [8] Seguro Social de Salud. “EsSalud”. In: *Norma Técnica del Programa Reforma de Vida. Sumak Kawsay: Vivir en Armonía, promovida por el Seguro Social de Salud. Lima* (2016).
- [9] CONVENIO TIARCO ENTRE EL SEGURO SOCIAL and DE SALUD ES-SALUD Y LA. “á# EssaLud”. In: (2014).
- [10] M. A. Valdez, A. D. Castillo, and E. P. Flores. “Aprendizaje automático para predicción de anemia en niños menores de 5 años mediante el análisis de su estado de nutrición usando minería de datos”. In: *Revista de Salud Pública* 19.2 (2023), pp. 128–138.
- [11] Miguel Ángel Rulo Zevallos Peñalva. “Estilo de liderazgo situacional y clima organizacional en el personal de enfermería Hospital III Yanahuara–EsSalud Arequipa-2016”. In: (2018).